Discovery and characterization of mammalian endogenous parvoviruses

Journal of Virology

Discovery and Characterization of Mammalian Endogenous Parvoviruses

Amit Kapoor, Peter Simmonds and W. Ian Lipkin
J. Virol. **2010, 84(24):12628. DOI: 10.1128/JVI.01732-10.**
Published Ahead of Print 13 October 2010.

JOURNAL OF VIROLOGY, Dec. 2010, p. 12628–12635 Vol. 84, No. 24
0022-538X/10/$12.00 doi:10.1128/JVI.01732-10

Discovery and Characterization of Mammalian Endogenous Parvoviruses▽†

Amit Kapoor,[1]* Peter Simmonds,[2] and W. Ian Lipkin[1]

Center for Infection and Immunity, Columbia University, New York, New York 10032,[1] and University of Edinburgh, Edinburgh, Scotland, United Kingdom[2]

Received 16 August 2010/Accepted 5 October 2010

Public databases of nucleotide sequences contain exponentially increasing amounts of sequence data from mammalian genomes. Through the use of large-scale bioinformatic screening for sequences homologous to exogenous mammalian viruses, we found several sequences related to human and animal parvoviruses (PVs) in the *Parvovirus* and *Dependovirus* genera within genomes of several mammals, including rats, wallabies, opossums, guinea pigs, hedgehogs, African elephants, and European rabbits. However, phylogenetic analysis of these endogenous parvovirus (EnPV) sequences demonstrated substantial genetic divergence from exogenous mammalian PVs characterized to date. Entire nonstructural and capsid gene sequences of a novel EnPV were amplified and genetically characterized from rat (*Rattus norvegicus*) genomic DNA. Rat EnPV sequences were most closely related to members of the genus *Parvovirus*, with >70% and 65% amino acid identities to nonstructural and capsid proteins of canine parvovirus, respectively. Integration of EnPV into chromosome 5 of rats was confirmed by PCR cloning and sequence analysis of the viral and chromosomal junctions. Using inverse PCR, we determined that the rat genome contains a single copy of rat EnPV. Considering mammalian phylogeny, we estimate that EnPV integrated into the rat genome less than 30 million years ago. Comparative phylogenetic analysis done using all known representative exogenous parvovirus (ExPV) and EnPV sequences showed two major genetic groups of EnPVs, one genetically more similar to genus *Parvovirus* and the other genetically more similar to the genus *Dependovirus*. The full extent of the genetic diversity of parvoviruses that have undergone endogenization during evolution of mammals and other vertebrates will be recognized only once complete genomic sequences from a wider range of classes, orders, and species of animals become available.

Approximately 8% of the human genome comprises endogenous retroviruses (ERV) (10, 15–17, 21). Molecular characterization of ERV has provided insights into the origin and evolution of their exogenous counterparts and also of their hosts (9, 10, 16, 17). While ERV are classically regarded as "junk" or "selfish" DNA, the maintenance of gene expression of some integrated ERV proviral sequences led to the proposal that ERV may be co-opted in certain cellular functional roles. For example, expression of *gag*- and *env*-encoded proteins of human ERV K (HERV-K) in the placenta may play roles in preventing rejection of the fetus (25). Until the recent discovery of endogenous Borna disease virus- and filovirus-like sequences in mammalian genomes (5, 15, 33), ERV were the only endogenous viruses known. The process of endogenization, whereby viral sequences are integrated in the genomes of their hosts (12), occurs when viral nucleic acid as DNA or cDNA integrates into chromosomes of reproductive germ line cells. In contrast to integration of reverse-transcribed proviral sequences of retroviruses, which is part of the retrovirus replicative cycle, integration of viral sequences of other viruses is cell mediated and serendipitous and creates embedded, fossilized genomic elements, incapable of generating infectious virions (5, 12, 15).

Exogenous parvoviruses (ExPVs) are ubiquitous and can cause a broad spectrum of diseases in animals, including enteritis, panleukopenia, hepatitis, erythrocyte aplasia, immune complex-mediated vasculitis, reproductive failure, and cerebellar ataxia (11). Vaccines against animal parvovirus infections are widely employed (2, 14, 27, 28, 31). The family *Parvoviridae* as currently defined comprises two subfamilies, *Densovirinae* and *Parvovirinae*, that infect nonvertebrate and vertebrate hosts, respectively (11). The International Committee on Taxonomy of Viruses (ICTV) has further classified the subfamily *Parvovirinae* into five genera primarily based on phylogenetic analysis: *Dependovirus*, *Bocavirus*, *Erythrovirus*, *Parvovirus*, and *Amdovirus*.

Members of the subfamily *Parvovirinae* are small, nonenveloped icosahedral viruses with single-stranded linear DNA genomes that frequently infect animals through the fecal-oral route (11). The genomes of most parvoviruses are nearly 5,000 nucleotides (nt) in length and comprise two transcriptional units, one encoding the capsid proteins and the other the nonstructural proteins. Protein-coding sequences are flanked on each side by noncoding palindromic repeats, also known as inverted terminal repeat (ITR) sequences, that play an important role in viral DNA replication (22, 23). Based on their replication requirements, parvoviruses can be classified as either autonomous parvoviruses or dependoviruses; the latter require external factors for replication (6). The most extensively studied dependoviruses are the adeno-associated viruses (AAV) that are used as gene therapy vectors. During their replication in the nuclei of infected host cells, the genomes of wild-type AAV integrate in a site-specific manner (chromo-

* Corresponding author. Mailing address: Center for Infection and Immunity, Columbia University, 722 West 168th Street, New York, NY 10032. Phone: (212) 304-5690. Fax: (212) 342-9044. E-mail: ak3117@columbia.edu.

† Supplemental material for this article may be found at http://jvi.asm.org/.

▽ Published ahead of print on 13 October 2010.

TABLE 1. Results of sequence similarity-based search for EnPVs in NCBI sequence databases[a]

Protein	Accession numbers (source) from:			
	Reference genomic sequence database (refseq_genomic)	NCBI Genomes/Chromosomes	Expressed sequence tags (est database)	Whole-genome shotgun reads (wgs database)
NS	NC_005101 (rat, ch-2), AC_000081 (rat, ch-13), NC_000069 (mouse, ch-3), NC_008806 (opossum, ch-6), NC_008803 (opossum, ch-3), NC_013678 (rabbit, ch-10), NC_009175 (horse, ch-X)	NC_005104 (rat, ch-5), NC_008803 (opossum, ch-3), NC_008806 (opossum, ch-6), NC_008808 (opossum, ch-8), C_013678 (rabbit, ch-10)	DW308476 (rat liver), B577555 (drosophila), C348320 (opossum liver), X610113 (pea aphid), V837559 (aphid), C387312 (pea aphid)	AABR05124553 (*Rattus norvegicus*), BQO010318785 (*Macropus eugenii*), BQO010599350 (*Macropus eugenii*), BQO010031620 (*Macropus eugenii*), AKN02032906 (*Cavia porcellus*), AKN02030352 (*Cavia porcellus*), AIY01487966 (*Echinops telfairi*), AGU03013549 (*Loxodonta africana*), APE01526173 (*Myotis lucifugus*), BRN01283281 (*Tursiops truncatus*), AAE01015016 (*Tetraodon nigroviridis*), AAYZ01294085 (*Ochotona princeps*), BRP01170809 (*Pteropus vampyrus*)
VP1	NC_005104 (rat, ch-5), C_000069 (mouse, ch-3), C_008806 (opossum, ch-6), C_008808 (opossum, ch-8), C_008803 (opossum, ch-3), C_013678 (rabbit, ch-10)	NC_005104 (rat, ch-5), C_008806 (opossum, ch-6), C_008808 (opossum, ch-8), C_008803 (opossum, ch-3)	DW382746 (rat liver), W313020 (rat liver), Y609048 (opossum tissues), O892248 (bovine brain), O888893 (bovine brain)	ABQO010519946 (*Macropus eugenii*), ABQO010334457 (*Macropus eugenii*), ABQO010193462 (*Macropus eugenii*), ABQO010585939 (*Macropus eugenii*), AKN02030352 (*Cavia porcellus*), AGV020719236 (*Dasypus novemcinctus*)

[a] All BLAST searches were performed using tblastn criteria described in Materials and Methods. ExPVs used to perform tblastn searches are as follows (accession numbers of NS and VP1 proteins, respectively, are in parentheses): canine parvovirus (NP_041399 and NP_041400), Aleutian mink disease virus (NP_042872 and NP_042875), adeno-associated virus 1 (NP_049541 and NP_049542), human bocavirus 1 (YP_338086 and YP_338088), and human parvovirus B19 (NP_050019 and NP_050020). Accession numbers of the sequences that showed significant similarity to NS or VP proteins of parvoviruses and respective hosts of origin are shown. ch-2, chromosome 2.

some 19) (29, 37), resulting in latent infection of host cells. Several recent studies detected the presence of AAV genomes in tissues of humans and nonhuman primates as integrated virus in the host genome and/or in episomal closed circular form (8, 13, 30). As part of their replication cycle, all parvoviruses must enter the nuclei of their host cells and generate a double-stranded monomer replicative form. Some animal parvoviruses are known to cause persistent infections and long-term shedding (3, 4). Moreover, all parvovirus genomes have inverted terminal repeats that may facilitate their integration in host DNA (36).

While investigating the NCBI genomic sequence database for virus-like sequences, we observed sequences distantly related to canine parvovirus (CPV) integrated into the rat genome. Further homology searches revealed the presence of parvovirus-like sequences in several additional mammalian species. This work describes the first identification of endogenous parvoviruses (EnPVs) in different mammalian species, confirms their integration into the mammalian genome, and provides their preliminary phylogenetic classification.

MATERIALS AND METHODS

Identification of endogenous parvoviruses. We used one reference genome sequence representing one species for each of the five genera included in subfamily *Parvovirinae* to identify genetically related sequences in different NCBI sequence databases (Table 1). NCBI databases used to search for parvovirus related sequences included refseq_genomic (genomic entries from NCBI's Reference Sequence project), NCBI Genomes/Chromosomes (a database with complete genomes and chromosomes from the NCBI Reference Sequence project), est (a database of GenBank, EMBL, and DDBJ sequences from expressed sequence tags), and wgs (a database for whole-genome shotgun sequence entries). Default search criteria of NCBI tblastn (protein query search against a translated nonredundant database) were used, except for changing the highest expect score ($E = 10^{-10}$) to make the similarity search highly stringent. All exogenous parvovirus (ExPV) sequences were excluded from the analysis. The animal genomic sequences that showed significant tblastn similarity ($E < 10^{-10}$) to one or more parvovirus proteins were considered candidate EnPVs. These candidate EnPV sequences were used to research the database for other homologous sequences that would have been missed during similarity-based search using ExPVs.

Sequence acquisition and phylogenetic analysis of EnPV. After initial identification of potential EnPV sequences, genomic DNA of rats was used to test for the presence of EnPV (Fig. 1C). An outline of experiments to confirm the existence, integration, and genomic organization of rat EnPV is shown in Fig. S1 in the supplemental material. Rat DNA was obtained from the *Rattus norvegicus* lung epithelial cell line L2 (ATCC catalog number CCL-149) and from *Rattus norvegicus* kidney, liver, and brain using Trizol (Invitrogen). EnPV-specific PCR to detect the presence of rat EnPV in different samples targeted the VP gene. Briefly, primers EnPV-ratF1 (5′-ATGGCACCTCCGGCGAAAAG-3′) and EnPV-ratR1 (5′-CCTGGTCCCAGGTACTTGTAGCC-3′) were used in the first round of nested PCR and EnPV-ratF2 (5′-GCGAAAAGAGCCAGGAGAGGTAA-3′) and EnPV-ratR2 (5′-CAGGTACTTGTAGCCCGGAGG-3′) were used the in second round. For the first round of nested PCR, 2.5 μl of each specimen DNA was mixed with 5.2 μl 10× polymerase reaction buffer (Qiagen), 1.2 μl each deoxynucleoside triphosphate (dNTP) (10 mM), 20 pmol forward (EnPV-ratF1) and reverse (EnPV-ratR1) primers, 0.5 μl HotStart *Taq* DNA polymerase (Qiagen), and 33.5 μl diethyl pyrocarbonate (DEPC)-treated water, in a total reaction volume of 50 μl. The reaction was performed using initial denaturation at 95°C for 7 min, followed by six cycles of 95°C for 40 s, 62°C for 45 s, and 72°C for 30 s, then 35 cycles of 95°C for 30 s, 59°C for 30 s, and 72°C for 30 s and a final extension at 72°C for 5 min. For the second round of nested PCR, identical cycling conditions were used, with an annealing temperature of 63°C for the first six cycles and annealing temperature of 60°C for the remaining 35 cycles. The reaction mixture for the second round included 0.5 μl PCR product from the first round. Products were visualized following electrophoresis

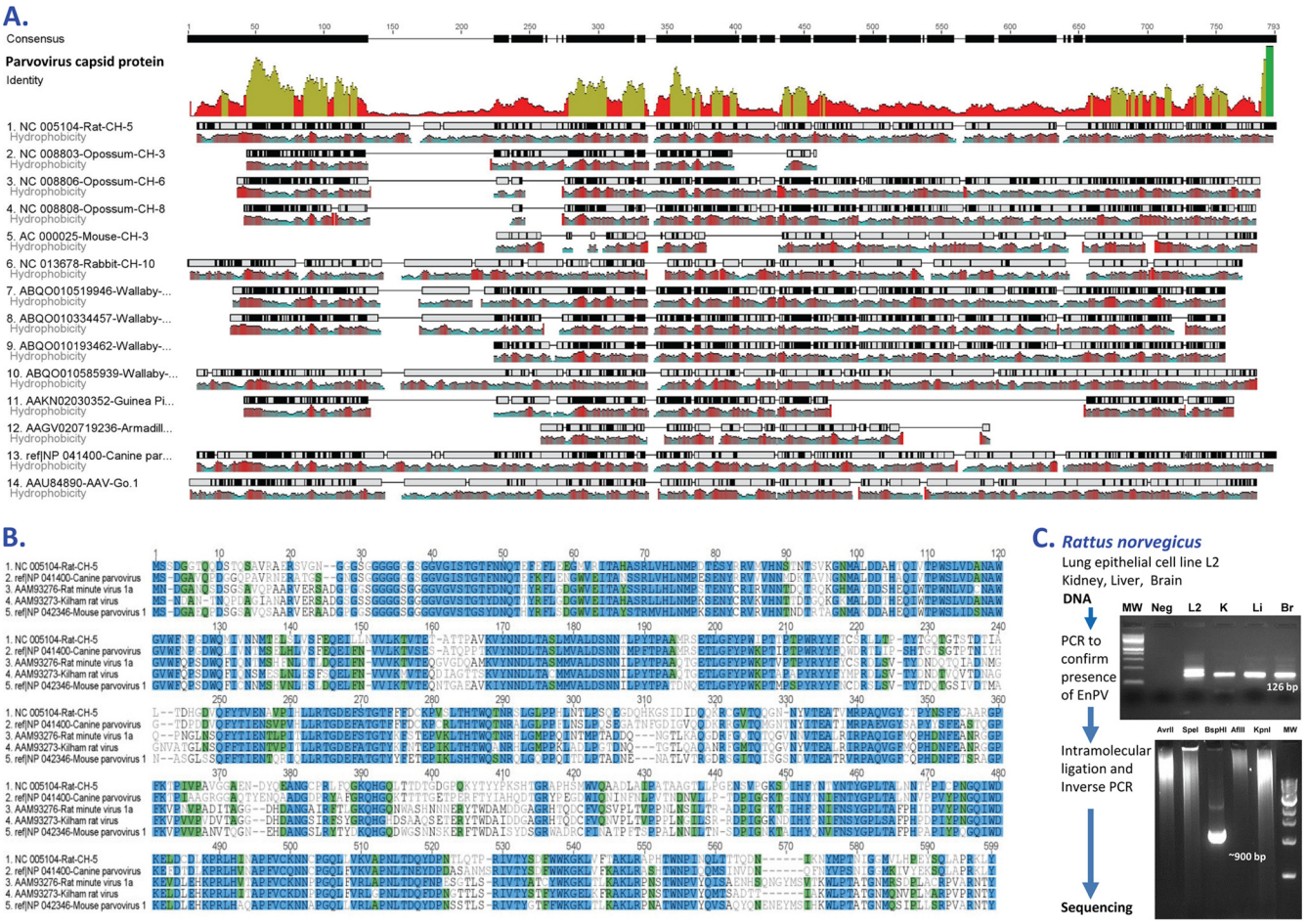

FIG. 1. (A) Genetic map of EnPV sequences encoding proteins with significant similarity to the structural proteins of canine parvovirus (CPV) and adeno-associated virus (AAV). Identity of amino acid residues and hydrophobicity scores of each sequence are plotted (sliding window size of 5 amino acids) to show the relatedness between structural proteins of mammalian EnPV and ExPV. Blocks in the protein alignment represent aligned amino acid residues, and lines represent gaps or missing data. In the identity plot, green represents amino acid residues conserved in >50% of sequences. In the hydrophobicity plot, red indicates higher concentrations of hydrophobic amino acids. Chromosome numbers for EnPV sequences of wallabies, guinea pigs, and armadillos are not known. (B) Protein alignment of the complete structural protein of rat EnPV with the other members of the genus *Parvovirus*. Amino acid residues identical in all sequences are shown in blue, amino acid residues identical in >50% sequences are shown in green, and variable residues are shown in a white background. (C) Results of specific PCR and inverse PCR for rat EnPV. Approximate sizes of amplification products are shown on the gel. MW, molecular weight marker; Neg, nontemplate control; K, kidney; Li, liver; Br, brain. L2 is a lung epithelial cell line, AvrII, SpeI, BspHI, AflII, and KpnI are restriction enzymes used for DNA digestion.

on 2% agarose gel. PCR products showing positive bands of approximately 126 bp (Fig. 1C), corresponding to the VP gene fragment of rat EnPV, were purified using a PCR purification kit (Qiagen) and directly sequenced from both directions. To acquire the complete genome of rat EnPV, PCR primers were designed to amplify five different overlapping fragments of genome using nested PCR conditions (see Table S1 in the supplemental material). All reactions for genomic PCR were performed in PCR buffer suitable for amplification of GC-rich genomic sequences as specified by the manufacturer (Takara; catalog no. RR02AG). All PCRs for junction and genomic PCRs used conditions similar to those described for inverse PCR below, except that the annealing temperatures for PCRs were 6°C below the average melting temperatures of the two primers. PCR products were sequenced from both directions. All EnPV sequences generated by PCR and extracted from the NCBI database were translated to proteins *in silico* for phylogenetic analysis.

Integration site analysis and copy number. To confirm the integration of EnPV in the host genome, two different approaches were used (see Fig. S1 in the supplemental material). First, the sequence information available in the NCBI database was used to design the primer pairs targeting the junction of the viral gene and host chromosome. Amplification and sequencing of the junctional region (junctional PCR) were then pursued to define the integration site sequence (see Fig. S1 and Table S1 in the supplemental material). Second, to independently confirm the integration site sequence and to determine the copy number of EnPV per host genome, we used an inverse PCR approach (Fig. 1C; see Fig. S1 in the supplemental material). The genomic DNA of a rat was digested with multiple enzymes, rejoined under conditions designed to favor intramolecular ligation, and employed as the template for PCR amplification (Fig. 1C; see Fig. S1 in the supplemental material). Restriction enzymes used were AvrII, SpeI, BspHI, AflII, and KpnI. Restriction enzyme-digested PCR products were self-ligated in reaction conditions favoring intramolecular ligation (15). The first round of inverse PCR used an inverse forward primer targeting the rat EnPV NS gene (INP-NS-F1, 5′-T*T*C*ACAGCCAAACTCAGGGCTCCA CATAC-3′) and an inverse reverse primer targeting the rat EnPV VP gene (INP-VP-R1, 5′-C*T*G*TTCCCTGGTCCCAGGTACTTGTA-3′). Both first-round primers were made exonuclease resistant through phosphorothioation (*) of the first three bases at the 5′ end. Similarly, the second round of inverse PCR used primers INP-NS-F2 (5′-CTCAGGGCTCCACATACATGG-3′) and INP-VP-R2 (5′-CCTGGTCCCAGGTACTTGTAGCC-3′). For the first round of nested PCR, 5 μl of restriction enzyme-digested and circularized DNA templates was mixed with 25 μl GC Buffer I (Takara; catalog no. RR02AG), 6 μl dNTP solution (10 mM), 25 pmol forward and reverse primers, 0.5 μl LA *Taq* DNA polymerase (Takara; catalog no. RR002A), and 11.5 μl DEPC-treated water, in a total reaction volume of 50 μl. The reaction was performed using initial

denaturation at 95°C for 8 min, followed by six cycles of 95°C for 40 s, 64°C for 45 s, and 72°C for 5 min, then 35 cycles of 95°C for 30 s, 59°C for 30 s, and 72°C for 5 min and a final extension at 72°C for 15 min. The reaction mixture for the second round included 0.5 μl PCR product from the first round. For the second round of nested PCR, the second-round primers were used under identical cycling conditions, with an annealing temperature of 66°C for the first six cycles and annealing temperature of 62°C for the remaining 35 cycles. Products were visualized following electrophoresis on 1.5% agarose gel (Fig. 1C) and sequenced from both ends.

EnPV expression analysis. Total nucleic acids were extracted from all tissue samples of rats using Trizol (Invitrogen). To determine the presence of EnPV-derived transcript RNA, total nucleic acid was digested with DNase followed by PCR targeting the amino terminus of the VP protein of the rat EnPV. Amplification products were visualized after size fractionation by electrophoresis in 1.5% agarose gels. All PCRs included a non-DNase-treated positive control to confirm successful amplification of the targeted EnPV sequence in each PCR.

RESULTS

Identification of endogenous parvoviruses. The sequence of the structural protein (VP) of canine parvovirus (NC_041400) was used as the template in a tblastn search to identify genetically related sequences in the NCBI database. Our first sequence similarity search results indicated the presence of parvovirus VP-like sequences in several mammalian species (Table 1). All the genomic sequences were extracted from the database, conceptually translated, and aligned to VP protein sequences of exogenous parvoviruses (ExPV) for identifying the different fragments of the VP gene in each EnPV (Fig. 1A). Results of our analysis suggested that rat (*Rattus norvegicus*) chromosome 5 (NC_005104) contained the coding sequence for almost the entire structural protein and that this sequence was similar to those encoding VP proteins of different members of genus *Parvovirus* (Fig. 1B).

Characterization of EnPV in rats (rat EnPV in chromosome 5). The tblastn search using canine parvovirus (NC_001539) nonstructural (NS) protein against the NCBI Genomes/Chromosomes database showed the presence of a highly similar sequence in *Rattus norvegicus* chromosome 5. The translated sequence comprised 281 amino acids (aa) with >70% identity to the N-terminal half of canine parvovirus NS proteins and an expected score of $2e^{-98}$. To determine the location of the rat EnPV nonstructural protein coding sequence in chromosome 5, the corresponding nucleotide sequence was extracted and used to perform a BLAT search against databases available at the University of California Santa Cruz Web server (http://genome.ucsc.edu/) (20). DNA BLAT works by keeping an index of the entire genome in memory and is designed to quickly find sequences of 95% and greater similarity containing 25 bases or more (19). The BLAT results showed the location of rat EnPV integration between chromosomal loci 5q22 and 5q24, nearly in the middle of chromosome 5 (Fig. 2A). Nucleotide alignment of the homologous chromosome in mice, humans, and dogs showed this region to be nonconserved among these species. Detailed analysis of the neighboring sequence resulted in characterization of three other large open reading frames (ORFs), whose products showed highly significant similarity to virion proteins (VP) of canine parvovirus (Fig. 2A). The NS and VP coding sequences of rat EnPV were found in close proximity (within 4,000 bp) on chromosome 5. However, in contrast to the genomic organization of known ExPVs, the NS protein ORF of rat EnPV was in reverse orientation with respect to the capsid ORFs (Fig. 2A). Phylogenetic analysis of the 281-aa rat EnPV NS sequence showed the coding sequence to be equidistant from the NS coding sequences of canine, porcine, mouse, and rat ExPVs. The NS protein of rat EnPV showed >70% amino acid identity to the NS proteins of different parvoviruses (Fig. 2B). Protein sequences encoded by three different ORFs of rat EnPVs related to the parvovirus VP protein were aligned to known ExPV proteins and combined to deduce a 744-aa protein representing the product of the complete structural gene ORF (Fig. 1B). Phylogenetic analysis of this 744-aa protein showed that the full-length VP protein of rat EnPV was most closely related to the capsid protein of canine parvovirus, with >65% protein identity (Fig. 1B and 2B).

The presence of rat EnPV sequences and their sites of integration in *Rattus norvegicus* chromosome 5 were confirmed by three different specific assays (Fig. 1C; see Fig. S1 in the supplemental material). In one assay the complete nucleotide sequence of rat EnPV in chromosome 5 was generated by overlapping PCR followed by sequencing. Our results showed complete concordance with the NCBI chromosome sequence data, including the presence of stop codons and frameshifts needed to generate a different ORF for the structural protein (Fig. 2A). In a second assay, the junctions between the EnPV and chromosomal neighboring sequence were amplified by PCR and sequenced (see Fig. S1 in the supplemental material). Analysis of 400 nt of chromosomal sequence on either side of the integrated EnPV confirmed the fidelity of the NCBI database and led to identification of long interspersed nuclear elements (LINE) of the L1MD family near the 3′ terminal end of rat EnPV in chromosome 5 (Fig. 2A). We did not find any sequence similarity between the LINE-like repeat element of rat EnPV and the ITRs of closely related members of the genus *Parvovirus*. A specific inverse PCR assay targeting the distal ends of rat EnPV in chromosome 5 showed the presence of a single prominent amplified product, indicating the presence of only one integrated rat EnPV copy per genome (Fig. 1C). The inverse PCR product was sequenced to further confirm the integration of rat EnPV in chromosome 5.

Identification and distribution of EnPV in other host species. To investigate the existence of EnPV in other species, we used ExPVs representing each of the five genera included in the subfamily *Parvovirinae* ExPVs as the query to perform a tblastn search against different NCBI sequence databases (Table 1). Sequences with significant similarity, defined as BLAST E values of $<10^{-10}$ versus one or more parvovirus proteins, were considered EnPV. EnPV sequences were detected in genomes of several mammalian species (Table 1), including rats, opossums, wallabies, guinea pigs, hedgehogs, armadillos, American pikas, elephants, rabbits, and horses, as well as in the genomes of puffer fish and some invertebrates, including drosophila and pea aphids. Genomic sequences with similarity to coding sequences of NS proteins of parvoviruses were more numerous than those with similarity to coding sequences of VP proteins. This finding was anticipated because NS motifs essential for viral DNA replication are likely to be more conserved during evolution. Moreover, although these single-stranded viruses have a DNA genome and use cellular replication machinery, their rate of nucleotide substitution is closer to that of RNA viruses than to that of double-stranded DNA viruses (32).

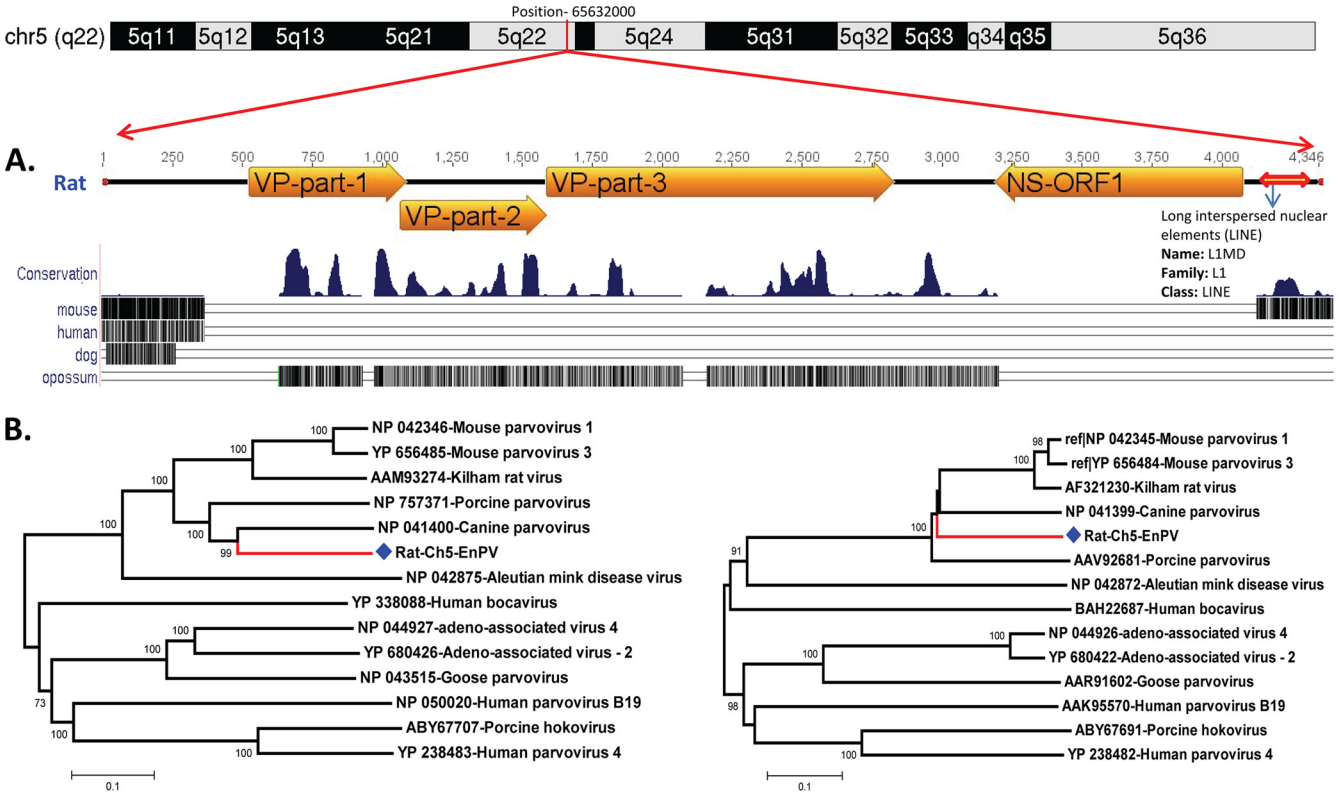

FIG. 2. Genomic organization and phylogenetic relationship of rat EnPV. (A) Map of chromosome 5 of rats (*Rattus norvegicus*) showing integration site of EnPV, open reading frame (ORF) length, and orientation (arrowheads of yellow boxes) and organization of capsid (VP) and nonstructural (NS) protein genes. The LINE-like element identified at the end of NS ORF is shown as a double-headed arrow. Sequence diversity plots for corresponding chromosome regions in genomes of mice, humans, dogs, and opossums showed the absence of rat EnPV sequences in other animals. (B) VP and NS protein sequences of known exogenous parvovirus reference strains available in GenBank were used to determine the genetic relatedness of rat EnPVs by neighbor-joining analysis of pairwise distances between translated amino acid sequences. Accession numbers of sequences used precede the names of corresponding parvovirus species. Bootstrap resampling was used to determine robustness of individual clades (values above 70% are shown above the branches).

EnPV sequences related to VP protein were more common in various rodents and marsupials than in other mammalian groups like primates, canines, and ruminants (Table 1). Whereas in mice, rats, rabbits, and armadillos only one EnPV sequence was detected, multiple diverse EnPV sequences were found in opossums, wallabies, and guinea pigs. Although mice and rats are the most closely related species in which EnPVs were found, the mouse EnPV was genetically closest to members of *Dependovirus* genus, in contrast to the rat EnPV (genetically closest to members of *Parvovirus* genus).

Phylogeny of mammalian EnPV. To determine the phylogenetic relationships of EnPV sequences to exogenous parvoviral sequences, all genomic sequences were translated *in silico* and joined manually using reference protein sequences of the most closely related known ExPV. All protein alignments were generated using the ClustalW program implemented in MEGA4.1. To accommodate the missing partial protein sequences of EnPV and the presence of stop codons and high genetic diversity between sequences, we used the BLOSUM protein weight matrix after reducing the gap opening penalty to 5 and gap extension penalty to 0.1 for both pairwise and multiple sequence alignments. Phylogenetic analysis done using default criteria also resulted in a tree topology similar to that generated using the modified criteria mentioned above. Minimal manual editing of EnPV sequences was performed to ensure an unbiased analysis. Although many of the protein sequences are of different lengths in both the NS and VP protein alignments, it is now well known that sequences of different lengths can be accurately placed in phylogenetic trees (35).

On the basis of their genetic relatedness to parvoviruses, the EnPV sequences extracted from the NCBI database were divided into two groups. The first group included the genomic sequences that showed significant similarity to VP genes of parvoviruses, and the second group included genomic sequences that showed similarity to NS genes of parvoviruses (Table 1). EnPV sequences similar to VP genes of parvoviruses were further classified into two heterogeneous phylogenetic clusters: one genetically closer to members of genus *Parvovirus*, named EnPV group A (EnPV-A), and the other genetically closer to members of genus *Dependovirus*, named EnPV group B (Fig. 3A). EnPV-A included virus sequences derived from genomes of wallabies, guinea pigs, opossums, and rats. All the members of EnPV-A, excluding the rat EnPV, clustered together to form a distinct clade whose members were genetically closer to each other than to known ExPVs. The capsid protein of rat EnPV was most closely related to the capsid protein of canine parvovirus and showed 65% amino acid identity to the

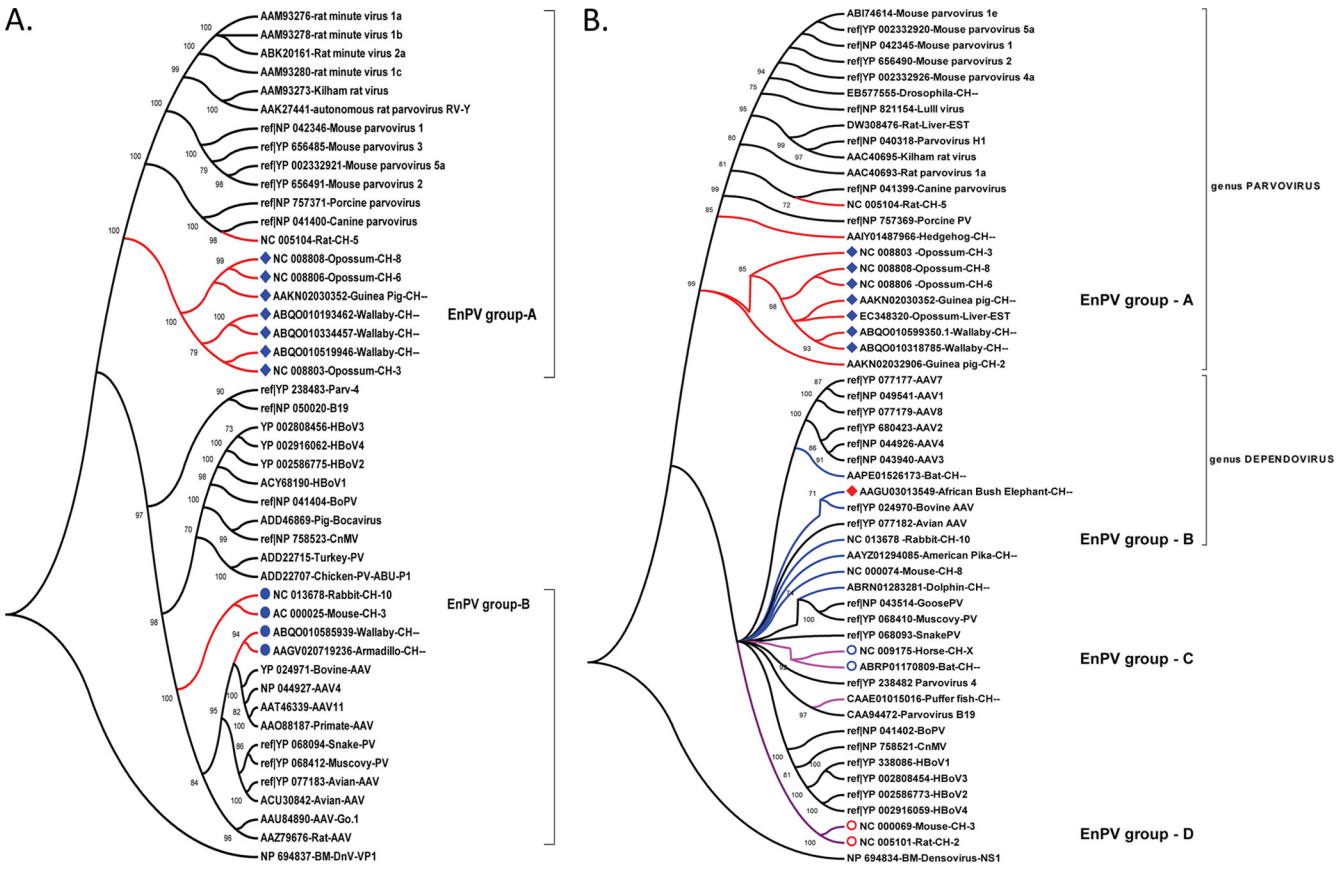

FIG. 3. Genetic diversity and phylogenetic analysis of the capsid and nonstructural proteins of EnPVs. Phylogenetic trees were constructed by neighbor-joining of pairwise protein distances based on alignments generated by ClustalW implemented in MEGA4.1. For clarity the accession numbers of sequences used precede the names of the corresponding parvovirus species. Bootstrap resampling was used to determine robustness of individual clades (values above 70% are shown above the branches). All trees are condensed to show the genetic relatedness not the distances. (A) VP gene sequences of different parvovirus species representing all genera of subfamily *Parvovirinae*, unclassified parvoviruses (human parvovirus 4 and turkey and chicken parvoviruses), and one species of *Densovirinae* were phylogenetically compared to the VP gene-like sequences of EnPVs from different animals. The two different genetic groups of EnPVs based on VP protein genetic relatedness are designated A and B (details in Results). (B) NS protein sequences of different parvovirus species representing all genera of subfamily *Parvovirinae*, unclassified parvoviruses (human parvovirus 4 and turkey and chicken parvoviruses), and one species of *Densovirinae* were compared with the NS protein-like sequences of EnPVs from different animals. The four different genetic groups of EnPVs based on NS protein genetic relatedness are designated A to D and are described in detail in Results. CH, chromosome (CH–, unknown chromosome number); EST, expressed sequence tags; BoPV, bovine parvovirus; CnMV, canine minute virus; HBoV, human bocavirus.

canine parvovirus protein. The genus *Parvovirus* includes ExPVs known to cause diseases in mouse, rat, porcine, and canine hosts. Intriguingly, the rat EnPV capsid protein was more closely related to those of canine ExPVs than to those of rat or mouse ExPVs (Fig. 1B and 2B).

The NS gene-related EnPV sequences could be classified into four heterogeneous clusters, tentatively named EnPV groups A to D (Fig. 3B). Two NS protein clusters, EnPV-A and -B, were similar to VP protein clusters described above and were genetically closest to members of genera *Parvovirus* and *Dependovirus*, respectively. EnPV sequences that clustered together with the NS genes of human parvovirus B19 and human parvovirus 4 are in EnPV-C and included EnPV sequences found in bats, horses, and puffer fish. A few EnPV NS protein sequences that were highly diverse and equidistant from those of all known ExPV genera formed an outlier group, which we designated EnPV group D. Like the phylogeny based on VP protein sequences, the phylogeny based on NS proteins of EnPV in rats, opossums, wallabies, and guinea pigs indicated genetic similarity to members of genus *Parvovirus* (EnPV-A). Although not conclusive, this finding is compatible with endogenization of both parvovirus genes together from the same exogenous virus.

Expression of EnPV in hosts. EnPV sequences were found in the NCBI expressed sequence tag (est) database, indicating that they were expressed at the mRNA level in host animals. Indeed, in rats we found expression of sequence tags representing two different genomic regions of rat EnPV (Table 1). To independently test for the expression of EnPV, we looked for the presence of viral RNA transcripts in rat tissues. Total nucleic acid was extracted, and a fraction of the sample was digested with RNase-free DNase to ensure amplification of only RNA. Reverse transcription-PCR (RT-PCR) assays targeting the structural gene of rats showed no amplification products. Additionally, no EnPV transcripts were found in actively replicating rat lung epithelial cells (rat L2 cells). How-

ever, results of these experiments cannot rule out the transient or more-tissue-restricted expression of EnPV in rats.

DISCUSSION

Our data indicate that parvoviruses are integrated into the genomes of a wide range of hosts, including rats, wallabies, opossums, guinea pigs, hedgehogs, African elephants, and European rabbits. The three other viruses known to be integrated into host genomes, retroviruses, Borna viruses (5, 15), and filoviruses (5, 33), have RNA genomes and require DNA replication intermediates for integration. Although parvoviruses have small single-stranded DNA genomes, their replication within the host cell nucleus requires synthesis of a double-stranded monomer replicative form that could facilitate integration into host genomes (6). Consistent with this model, several EnPVs characterized here showed close proximity in the integration sites of NS and VP gene sequences, suggesting endogenization of complete viral DNA genomes rather than independent RNA transcripts encoding different genes. However, we failed to detect the noncoding sequences of genetically related parvoviruses near capsid or nonstructural genes of EnPV found in rats. Interestingly the endogenization of EnPVs is likely to be more recent (<20 to 30 million years ago) than that of filoviruses, since the EnPVs from genetically related host species (rats and mice) are not orthologous like their host animals (1, 5, 33).

Endogenous viruses are fossilized exogenous viruses (5, 12). Over time host genetic mechanisms such as mutation and genomic rearrangements lead to the eventual degeneration of endogenous virus genes; attrition of their capacity to express viral gene products may indeed be a benefit to the host (5). Detailed analysis of rat EnPV genetic diversity and genomic organization suggests that rat EnPV may represent an early progenitor of the genus *Parvovirus*. The absence of rat EnPV-like sequences in the canine and porcine genomes suggests that either rats were the only natural host of this virus or endogenization of this virus in the genomes of the other related mammalian species was unsuccessful. All known parvoviruses have similar genomic organizations wherein the NS and VP genes are present in the same orientation. In rat EnPV the NS and VP genes are oriented oppositely in spite of being in close proximity, which suggests that, after the integration of an intact genome, small-scale, local genome rearrangements disrupted the integrated parvoviral sequence. The presence of a LINE-like repeat sequence near the truncated NS gene in rat EnPV suggests a role for these elements in integration of viral genes, similar to the role these elements play in generating cellular pseudogenes (5). However, the presence of both parvovirus NS and VP genes in close proximity in several EnPVs suggests that direct insertion of double-stranded viral DNA replication intermediates into the chromosome is also plausible.

Endogenous retroviruses may protect the host from infection and disease by similar exogenous viruses (7), a resistance mechanism that has also recently been proposed to underlie the integration of potyvirus and dicistrovirus genome sequences in insects (26). None of the endogenous viruses described thus far has a genetically identical exogenous counterpart (9, 10). Similarly, we cannot ascertain whether the presence of rat EnPV protects rats from infection by closely related exogenous viruses as no genetically similar exogenous virus has been discovered. However, we do know that the presence of rat EnPV does not provide protection against several known rat parvoviruses that naturally infect rats and whose VP and NS proteins are >25 to 35% different in amino acid sequence (Fig. 3). Although we did not detect expression of rat EnPV mRNA in tissue samples from rat organs or a continuous cell line, expression may be transient. We also found sequences of expressed sequence tags in the NCBI database, suggesting the expression of EnPV in rats, opossums, and aphids (Table 1).

The observed genetic diversity between NS (>25%) and VP (>35%) proteins of rat EnPV and known parvoviruses is consistent with faster evolution rate of the parvovirus VP protein by as much as 10%. Based on comparative phylogenetic analysis, most of the EnPVs are either genetically close to members of the genus *Parvovirus* or *Dependovirus* (24). In the opossum we observed that all three different EnPV VP coding sequences integrated in different chromosomes belong to EnPV-A. In the absence of a mechanism for expression, reverse transcription, and reintegration of viral sequences (as documented for ERV and LINE), this suggests multiple endogenization events over time. Indeed, in the wallaby, we found integrated VP sequences of both the EnPV-A and EnPV-B clusters. Unlike other endogenous viruses reported (5, 9, 15, 34), the EnPV-like sequences were not detected in genomes of primates and were more prevalent in small mammals and marsupials. The genetic diversity of the combined data set of EnPV sequences examined and lack of homologous integrated sequences in related rodent species suggest that multiple endogenization events occurred after the speciation of rodents 25 to 35 million years ago (1). The genetic relatedness between the EnPVs and their host species is also discordant, consistent with endogenization after host speciation. For example, the EnPVs of rats and mice are genetically more distant than EnPVs of mice and rabbits. Considering that rats and mice separated 20 to 40 million years ago (1), the endogenization time of rat EnPV in chromosome 5 appears to be more recent. Interestingly we found no EnPV genetically related to members of erythroviruses, avian parvoviruses, or bocaviruses, with the exception of the NS gene sequences of EnPVs found in puffer fish chromosomes that were genetically closest to parvovirus B19 (Fig. 3B). Bocaviruses are unique among parvoviruses in that they contain a middle ORF called NP1 between the NS and VP coding regions (18, 24). The tblastn search using the bocavirus NP1 protein did not detect any significantly related sequence in the NCBI database.

In summary, we report parvoviruses as the first small DNA viruses that exist as endogenous viruses integrated in the host chromosomes and demonstrate that they are widespread among several mammalian species. Our observations based on genetic diversity suggest that EnPVs were generated as results of multiple endogenization events. We believe the EnPVs are likely to be more widely dispersed than described here and that their complete diversity will be recognized only as genomic sequences of more mammalian species become available.

ACKNOWLEDGMENTS

We thank Brian Hjelle for helpful comments, Mady Hornig for critical reagents, and Natasha Mehta and Natasha Qaisar for technical assistance.

Our work was supported by National Institutes of Health grant awards (AI090196A, AI079231, AI57158, AI070411, and EY017404) and by an award from the Department of Defense.

REFERENCES

1. **Adkins, R. M., E. L. Gelke, D. Rowe, and R. L. Honeycutt.** 2001. Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. Mol. Biol. Evol. **18:**777–791.
2. **Appel, M. J.** 1999. Forty years of canine vaccination. Adv. Vet. Med. **41:** 309–324.
3. **Bass, L. R., and F. M. Hetrick.** 1978. Human lymphoblastoid cells as hosts for parvoviruses H-1 and rat virus. J. Virol. **25:**486–490.
4. **Bass, L. R., and F. M. Hetrick.** 1978. Persistent infection of a human lymphocyte cell line (Molt-4) with the Kilham rat virus. J. Infect. Dis. **137:**210–212.
5. **Belyi, V. A., A. J. Levine, and A. M. Skalka.** 2010. Unexpected inheritance: multiple integrations of ancient Bornavirus and Ebolavirus/Marburgvirus sequences in vertebrate genomes. PLoS Pathog. **6:**e1001030.
6. **Berns, K. I.** 1990. Parvovirus replication. Microbiol. Rev. **54:**316–329.
7. **Bishop, K. N., M. Bock, G. Towers, and J. P. Stoye.** 2001. Identification of the regions of Fv1 necessary for murine leukemia virus restriction. J. Virol. **75:**5182–5188.
8. **Chen, C. L., R. L. Jensen, B. C. Schnepp, M. J. Connell, R. Shell, T. J. Sferra, J. S. Bartlett, K. R. Clark, and P. R. Johnson.** 2005. Molecular characterization of adeno-associated viruses infecting children. J. Virol. **79:**14781–14792.
9. **Coffin, J. M., M. Champion, and F. Chabot.** 1978. Nucleotide sequence relationships between the genomes of an endogenous and an exogenous avian tumor virus. J. Virol. **28:**972–991.
10. **Coffin, J. M., P. N. Tsichlis, K. F. Conklin, A. Senior, and H. L. Robinson.** 1983. Genomes of endogenous and exogenous avian retroviruses. Virology **126:**51–72.
11. **Fauquet, C. M., M. A. Mayo, J. Maniloff, U. Desselberger, and L. A. Ball (ed.).** 2005. Virus taxonomy: the eighth report of the International Committee on Taxonomy of Viruses. Academic Press, San Diego, CA.
12. **Feschotte, C.** 2010. Virology: Bornavirus enters the genome. Nature **463:**39–40.
13. **Gao, G., L. H. Vandenberghe, M. R. Alvira, Y. Lu, R. Calcedo, X. Zhou, and J. M. Wilson.** 2004. Clades of adeno-associated viruses are widely disseminated in human tissues. J. Virol. **78:**6381–6388.
14. **Hoelzer, K., and C. R. Parrish.** 2010. The emergence of parvoviruses of carnivores. Vet. Res. **41:**39.
15. **Horie, M., T. Honda, Y. Suzuki, Y. Kobayashi, T. Daito, T. Oshida, K. Ikuta, P. Jern, T. Gojobori, J. M. Coffin, and K. Tomonaga.** 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature **463:**84–87.
16. **Hughes, J. F., and J. M. Coffin.** 2004. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. Proc. Natl. Acad. Sci. U. S. A. **101:**1668–1672.
17. **Hughes, J. F., and J. M. Coffin.** 2002. A novel endogenous retrovirus-related element in the human genome resembles a DNA transposon: evidence for an evolutionary link? Genomics **80:**453–455.
18. **Kapoor, A., E. Slikas, P. Simmonds, T. Chieochansin, A. Naeem, S. Shaukat, M. M. Alam, S. Sharif, M. Angez, S. Zaidi, and E. Delwart.** 2009. A newly identified bocavirus species in human stool. J. Infect. Dis. **199:**196–200.
19. **Kent, W. J.** 2002. BLAT—the BLAST-like alignment tool. Genome Res. **12:**656–664.
20. **Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler.** 2002. The human genome browser at UCSC. Genome Res. **12:**996–1006.
21. **Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al.** 2001. Initial sequencing and analysis of the human genome. Nature **409:**860–921.
22. **LeFebvre, R. B., and K. I. Berns.** 1984. Unique events in parvovirus replication. Microbiol. Sci. **1:**163–167.
23. **LeFebvre, R. B., S. Riva, and K. I. Berns.** 1984. Conformation takes precedence over sequence in adeno-associated virus DNA replication. Mol. Cell. Biol. **4:**1416–1419.
24. **Lukashov, V. V., and J. Goudsmit.** 2001. Evolutionary relationships among parvoviruses: virus-host coevolution among autonomous primate parvoviruses and links between adeno-associated and avian parvoviruses. J. Virol. **75:**2729–2740.
25. **Malassine, A., S. Blaise, K. Handschuh, H. Lalucque, A. Dupressoir, D. Evain-Brion, and T. Heidmann.** 2007. Expression of the fusogenic HERV-FRD Env glycoprotein (syncytin 2) in human placenta is restricted to villous cytotrophoblastic cells. Placenta **28:**185–191.
26. **Maori, E., S. Lavi, R. Mozes-Koch, Y. Gantman, Y. Peretz, O. Edelbaum, E. Tanne, and I. Sela.** 2007. Isolation and characterization of Israeli acute paralysis virus, a dicistrovirus affecting honeybees in Israel: evidence for diversity due to intra- and inter-species recombination. J. Gen. Virol. **88:** 3428–3438.
27. **Patel, J. R., and J. G. Heldens.** 2009. Review of companion animal viral diseases and immunoprophylaxis. Vaccine **27:**491–504.
28. **Paul, M. A., L. E. Carmichael, H. Childers, S. Cotter, A. Davidson, R. Ford, K. F. Hurley, J. A. Roth, R. D. Schultz, E. Thacker, and L. Welborn.** 2006. 2006 AAHA canine vaccine guidelines. J. Am. Anim. Hosp. Assoc. **42:**80–89.
29. **Samulski, R. J., X. Zhu, X. Xiao, J. D. Brook, D. E. Housman, N. Epstein, and L. A. Hunter.** 1991. Targeted integration of adeno-associated virus (AAV) into human chromosome 19. EMBO J. **10:**3941–3950.
30. **Schnepp, B. C., R. L. Jensen, C. L. Chen, P. R. Johnson, and K. R. Clark.** 2005. Characterization of adeno-associated virus genomes isolated from human tissues. J. Virol. **79:**14793–14803.
31. **Schultz, R. D.** 2006. Duration of immunity for canine and feline vaccines: a review. Vet. Microbiol. **117:**75–79.
32. **Shackelton, L. A., C. R. Parrish, U. Truyen, and E. C. Holmes.** 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc. Natl. Acad. Sci. U. S. A. **102:**379–384.
33. **Taylor, D. J., R. W. Leach, and J. Bruenn.** 2010. Filoviruses are ancient and integrated into mammalian genomes. BMC Evol. Biol. **10:**193.
34. **Tomonaga, K., and J. M. Coffin.** 1998. Structure and distribution of endogenous nonecotropic murine leukemia viruses in wild mice. J. Virol. **72:**8289–8300.
35. **Wiens, J. J.** 2006. Missing data and the design of phylogenetic analyses. J. Biomed Inform. **39:**34–42.
36. **Yang, C. C., X. Xiao, X. Zhu, D. C. Ansardi, N. D. Epstein, M. R. Frey, A. G. Matera, and R. J. Samulski.** 1997. Cellular recombination pathways and viral terminal repeat hairpin structures are sufficient for adeno-associated virus integration in vivo and in vitro. J. Virol. **71:**9231–9247.
37. **Young, S. M., Jr., D. M. McCarty, N. Degtyareva, and R. J. Samulski.** 2000. Roles of adeno-associated virus Rep. protein and human chromosome 19 in site-specific recombination. J. Virol. **74:**3953–3966.